

Aix-Marseille Université
Master Sciences et Technologie
Mention Bioinformatique, Biochimie structurale et Génomique
Spécialité professionnelle Bioinformatique et Génomique

2015-2016

DÉVELOPPER UN SYSTÈME DE GESTION DE DONNÉES POUR LE SUIVI DES DISPOSITIFS EXPÉRIMENTAUX AGRONOMIQUES

BOULNEMOUR MEDHI

Encadrement du stage
Sandrine Auzoux, David Pot, Lauriane Rouan



Centre de coopération internationale en recherche
agronomique pour le développement (Cirad)
Avenue Agropolis
34398 Montpellier cedex 5

REMERCIEMENTS

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon stage.

Tout d'abord, j'adresse mes remerciements à Mme Nabila Yahiaoui chercheuse dans l'UMR Agap qui m'a permis de postuler dans ce centre de recherche, et de trouver ce stage qui était en totale adéquation avec mes attentes.

Je tiens à remercier vivement mes maitres de stage, Mme Sandrine Auzoux, informaticienne dans l'UR Aïda, Mme Lauriane Rouan chercheuse biostatisticienne dans l'UMR Agap et M. David Pot chercheur dans l'UMR Agap et responsable du projet BFF, pour leur accueil, le temps passé ensemble et le partage de leur expertise au quotidien. Grâce à leur confiance j'ai pu m'accomplir totalement dans mes activités. Je les remercie aussi pour leurs lectures et les corrections apportées à ce manuscrit.

Je remercie également toute l'équipe ID (Intégration des Données) de l'UMR Agap pour leur accueil, leur esprit d'équipe et leur aide précieuse qui a contribué à la réussite de ce projet.

Mes remerciements s'adressent aussi personnellement à tous les professeurs de ma discipline.

Je tiens aussi à remercier toute l'unité Agap pour leur agréable accueil, et toutes les personnes qui m'ont aidées et soutenues, de près comme de loin, dans la réalisation de ce travail.

Je remercie également Emilie et tous mes amis pour leur soutien.

Enfin, je tiens à dédier ce travail à mes chers parents ainsi qu'à toute ma famille.

TABLES DES MATIÈRES

Remerciements	3
Tables des matières	5
Glossaire	7
Présentation de l'organisme d'accueil.....	11
PRESENTATION DU CIRAD.....	11
LE PROJET BIOMASS FOR THE FUTURE.....	13
Introduction.....	15
DU GENOTYPE AU PHENOTYPE : CONTEXTE.....	15
INITIATIVES EXISTANTES.....	15
<i>Ecofi : la base de données expérimentale du cirad.....</i>	<i>17</i>
<i>Ephesis : le système d'information de l'urgj</i>	<i>17</i>
<i>Le breeding management system de l'ibp</i>	<i>19</i>
<i>Phenome networks et agrobase</i>	<i>21</i>
<i>Conclusion</i>	<i>21</i>
DAPHNE	21
<i>Le cahier des charges</i>	<i>23</i>
<i>Les caractéristiques du système d'informations.....</i>	<i>23</i>
Matériels et méthodes.....	25
VOLUMETRIE ET COMPLEXITE DES DONNEES.....	25
APPROCHE DE DEVELOPPEMENT	27
LA BASE DE DONNEES.....	27
<i>Présentation des packages et cas d'utilisations</i>	<i>27</i>
LE FRAMEWORK	31
MODELE-VUE-CONTROLEUR.....	33
<i>Le contrôleur.....</i>	<i>33</i>
<i>La vue</i>	<i>33</i>
<i>Le modèle</i>	<i>33</i>
SYTEME DE GESTION DE VERSIONS : GIT.....	35
Résultats et discussion	37
IMPLEMENTATION DE LA BASE DE DONNEES.....	37
IMPLEMENTATION DE L'APPLICATION WEB	41
<i>Fonctionnalités.....</i>	<i>41</i>
DIFFICULTES RENCONTREES	45
<i>Les objectifs.....</i>	<i>45</i>
<i>L'agilité.....</i>	<i>45</i>
<i>La généricité et les méta-données</i>	<i>45</i>
Conclusion et perspectives	49
References	51
Annexes	53
ANNEXE 1 : LES OUTILS DE DEVELOPPEMENTS	53
<i>La base de données.....</i>	<i>53</i>
<i>L'application web.....</i>	<i>53</i>
ANNEXE 2 : L'ARCHITECTURE 3 TIERS.....	54
ANNEXE 3 : SCHEMA UML DE DAPHNE.....	55
ANNEXE 4 : POSTER SCIENTIFIQUE DE DAPHNE	56

GLOSSAIRE

AGAP	Amélioration génétique et adaptation des plantes méditerranéennes et tropicales
Back-end	Partie du code qui est exécuté coté serveur, il comprend généralement la manipulation et l'interrogation de base de données. Dans notre cas le langage utilisé est exclusivement écrit en PHP.
Biomasse	Quantité de matière vivante contenue dans une unité déterminée de surface ou de volume de l'environnement.
Bloc	Ensemble de parcelles où toutes les modalités sont représentées une seule fois, le bloc est homogène et la répartition des modalités est aléatoire. Un bloc correspond à une répétition.
CIRAD	Centre de coopération internationale en recherche agronomique pour le développement
Cotylédon	Expansion latérale de l'embryon végétal qui, généralement, est charnue et contient les réserves nutritives nécessaires au premier développement de la plante.
Dispositif expérimental	Ensemble d'unités physiques présentant une disposition particulière et caractéristique en vue de mettre en évidence, ou non, l'effet d'une action sur la culture
Épandage	Technique agricole consistant à répandre divers produits sur des zones cultivées.
Framework	Ensemble cohérent de composants logiciels structurels, qui sert à créer les fondations ainsi que les grandes lignes de tout ou d'une partie d'un logiciel (architecture).
Front-end	Partie du code qui est exécuté sur la plate-forme de l'utilisateur, dans notre cas c'est le code exécuté par le navigateur web. Langages : HTML, JavaScript, CSS.
Full-stack	Développement d'une application dans son ensemble, c'est à dire back-end et front-end.
IHM	Interface homme-machine
Interopérabilité	Capacité que possède un système informatique à fonctionner avec d'autres produits ou systèmes informatiques, existants ou futurs, sans restriction d'accès ou de mise en œuvre.
ITK	L'itinéraire technique cultural représente l'ensemble des interventions menées sur une parcelle.
Mafor	Matière fertilisante d'origine résiduaire comprenant l'ensemble des déchets d'origine anthropique liés aux activités industrielles, domestiques, urbaines ou agricoles et qui peut être valorisé en tant qu'engrais ou amendement pour l'agriculture.

Métadonnée	Donnée servant à définir ou décrire une autre donnée.
Monocotylédone	Groupe d'espèces végétales présentant une plantule à un cotylédon.
Parcelle	Unité expérimentale de base, où, selon un protocole, est appliquée une
Placette	Partie, à l'intérieur d'une parcelle, pour éviter les effets de bordure, sur laquelle sont réalisées des observations et/ou mesures : notation de l'enherbement, pesée de cannes, échantillon de sol....
UMR	Unité mixte de recherche
UR	Unité de recherche

PRÉSENTATION DE L'ORGANISME D'ACCUEIL

PRÉSENTATION DU CIRAD

Le CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement) est un établissement public à caractère industriel et commercial créé en 1984. C'est un organisme français de recherche agronomique et de coopération internationale pour le développement durable des régions tropicales et méditerranéennes.

Cet organisme travaille en partenariat avec les pays du Sud. Il produit et transmet de nouvelles connaissances pour accompagner leur développement agricole et contribuer au débat sur les grands enjeux mondiaux de l'agronomie. Du local au global, le CIRAD fait avancer la science et l'agriculture au service de tous. Le CIRAD mène des activités de coopération avec plus de 90 pays. Il met à disposition de la communauté scientifique nationale et internationale, un dispositif important de recherche et de formation.

Le CIRAD, comprend aujourd'hui 1650 personnes dont 800 chercheurs qui travaillent avec les pays du Sud. Le CIRAD est composé de 3 départements scientifiques que sont Performance des systèmes de production et de transformation tropicaux (Persyst), Environnement et sociétés (ES) et Systèmes biologiques (Bios), dans lesquels sont répartis 34 unités de recherche dont l'unité mixte de recherche Agap (Amélioration génétique et adaptation des plantes méditerranéennes et tropicales) qui fait partie du département Bios. Ainsi que l'UR Aïda (Agroécologie et intensification durable des cultures annuelles) qui fait partie du département Persyst (Figure 1).

L'UR Aïda a pour objectif l'étude, la conception et la proposition de systèmes de cultures annuels (canne à sucre, cotonnier, riz...), répondant aux exigences de performances agronomiques, technologiques et environnementales.

L'UMR Agap œuvre en faveur de la création de matériel végétal adapté aux systèmes de productions impliquant les principales espèces végétales tropicales et méditerranéennes (riz, blé, sorgho, bananiers, canne à sucre, etc.). Elle est composée de douze équipes de recherche, dont :

- L'équipe PAM concentre ses activités sur l'écophysiologie des monocotylédones, principalement le riz, le sorgho et le palmier à huile, tout en y associant des recherches en histologie, biochimie, statistiques et informatique.

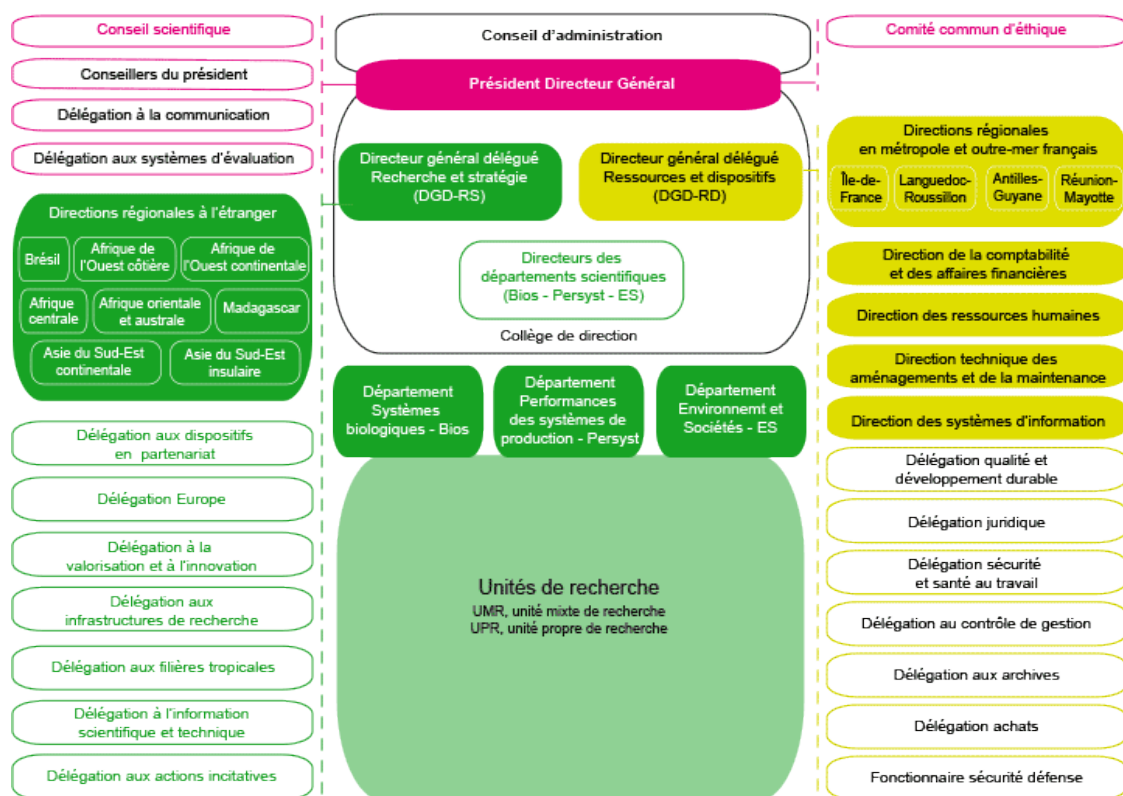


Figure 1 Organigramme du CIRAD

Elle contribue à l'optimisation de la performance des cultures et développe des connaissances et des modèles intégrés des processus morpho-physiologiques d'adaptation des plantes au milieu agro-écologique

- L'équipe d'intégration de données (ID). Elle se consacre au développement d'outils et de méthodes bio-informatiques originales qui s'appuient sur un grand nombre d'espèces et d'approches thématiques étudiés dans l'UMR. Le but est de permettre une analyse intégrée de l'adaptation des plantes cultivées tropicales et méditerranéennes aux facteurs environnementaux, divers et changeants. C'est dans l'équipe ID de l'unité Agap que j'ai été accueilli pour réaliser mon stage de fin d'étude.

LE PROJET BIOMASS FOR THE FUTURE

BFF (Biomass For the Future) est un projet de recherche scientifique multipartenaires, publiques et privés, qui s'étend sur la période 2012 à 2020. Il ambitionne de développer des systèmes de production et de valorisation de la biomasse issus de 2 espèces cibles :

- Le sorgho-fibre, une graminée résistante à la sécheresse annuelle à haut rendement adaptée à l'Europe méridionale.
- Le miscanthus, une graminée pérenne adaptée à des climats plus tempérés, avec un allongement de la saison de croissance et une dépendance moindre aux engrais et aux composés phytosanitaires que les espèces annuelles.

Ces cultures ont l'avantage de combiner un fort potentiel de production de biomasse, avec un impact minimal sur l'environnement. Pour répondre aux défis liés à la mise en place de filières de production et de valorisation de la biomasse, ce projet vise à optimiser les itinéraires techniques et développer des variétés dédiées aux valorisations ciblées en mobilisant les nouveaux outils de la génomique. Ensuite, le développement de petites et moyennes industries ancrées localement et fondées sur la biomasse doit servir de cas d'école. Cela permet de surmonter les obstacles technologiques, sociologiques et institutionnelles pour le développement futur à grande échelle de biomatériaux, des biocarburants de la deuxième génération et des industries de bioraffinage, qui à leur tour contribueront à consolider l'industrie locale. Cela doit créer un élan positif vers une adoption généralisée d'une bio-économie basée sur la connaissance en France.

INTRODUCTION

DU GÉNOTYPE AU PHÉNOTYPE : CONTEXTE

Le développement de variétés améliorées implique la compréhension fine du déterminisme génétique des caractères d'intérêts agronomiques et industriels. Cette compréhension permet en effet d'optimiser l'efficacité de la sélection variétale en optimisant l'identification des croisements à effectuer, en réduisant les coûts d'expérimentation (sélection basée sur le génotype et non sur le phénotype) et en accélérant les cycles de sélection. Néanmoins l'identification des facteurs génétiques contrôlant la variabilité des caractères d'intérêt nécessite la mise en relation de données hétérogènes.

Le phénotype (P) résulte des effets conjoints de 3 composantes : le génotype (G), l'environnement (E) et l'interaction entre le génotype et son environnement (G x E). Ceci est résumé par la formule suivante : $P = G + E + G \times E$

Puisque l'expression d'un gène n'est pas indépendante du milieu dans lequel ce gène s'exprime, il est important de pouvoir stocker les informations relatives à ces différentes composantes (informations génétiques, informations environnementales) et à leurs conséquences sur les phénotypes. Si aujourd'hui de nombreux systèmes pour stocker ou exploiter les ressources génétiques existent, c'est qu'on a compris l'importance de l'accès à ces informations. Cependant, nous sommes forcés de constater la difficulté de trouver un système d'information générique pour accueillir les données de ces interactions génotype-environnement en agronomie.

INITIATIVES EXISTANTES

Dans l'état actuel, chaque agronome/généticien/sélectionneur gère ses données de manière individuelle le plus souvent à l'aide de tableurs. La quantité, la diversité l'hétérogénéité des formats de stockage de ces données sont importantes. Cette méthode de gestion des données a montré ses limites sur plusieurs plans : l'accès, la qualité et la traçabilité des informations. Les données sont parfois redondantes et éparpillées sur plusieurs fichiers et plusieurs supports physiques. En conséquence, le temps passé à rechercher une information devient considérable.

Ces difficultés imposent le développement d'outils performants pour le stockage et l'exploitation des données. En outre la réalisation d'analyse intégrative est généralement rendue très difficile par la dispersion des données et leur absence de qualification. Suite à ce constat, les équipes du projet BFF ont décidé d'étudier les solutions existantes dans le domaine.

ECOFI : LA BASE DE DONNEES EXPERIMENTALE DU CIRAD

ECOFI est une base de données expérimentale développée par le CIRAD pour la gestion des essais en agroécologie. C'est une base de données relationnelle conçue à partir de l'analyse des jeux de données issus d'essais pluridisciplinaires sur la canne à sucre.

Ce système permet l'intégration des données phénotypiques de plusieurs espèces, à différentes échelles issues de plusieurs domaines (agronomie, entomologie, écophysiologie, malherbologie). Il utilise la technologie des métadonnées lui permettant d'ajouter librement autant de variables que l'on veut sans altérer la structure de la base de données. Dans les bases de données standards, chaque variable observée implique généralement une mise à jour du modèle de la base de données existante. Le modèle de données d'ECOFI ne nécessite pas une telle modification. La prise en compte d'une nouvelle variable est très facile et consiste seulement à l'ajout d'un nouvel enregistrement dans une table de métadonnées générique, après avoir mémorisé le nouveau libellé de la variable et sa définition (unité, type, échelle). Cette technologie réduit le nombre de tables, de colonnes et de cellules vides, et améliore les performances des requêtes de base de données (Auzoux Sandrine, 2016).

En revanche, ce système ne propose aucune gestion des échantillons, le dictionnaire des données d'ECOFI n'est pas en relation avec les ontologies existantes dans le domaine. Et il n'est pas capable de gérer plusieurs utilisateurs et partenaires. De plus, la description des dispositifs expérimentaux est limitée à 3 niveaux hiérarchiques, ce qui diminue grandement la capacité de l'outil à intégrer les dispositifs plus complexes.

EPHESIS : LE SYSTÈME D'INFORMATION DE L'URGI

L'URGI est une unité de recherche en génomique et bio-informatique de l'INRA, dédiée aux plantes et leurs pathogènes. Son activité de recherche porte sur l'intégration des données, la structure et la dynamique du génome. De plus, L'unité héberge une plate-forme bio-informatique française, Elixir.

C'est au travers de cette plate-forme que l'URGI est en train de développer Ephesis, un système d'information dédié à l'étude des interactions génotype-environnement. C'est le système principalement utilisé par les chercheurs et agronomes de l'INRA. Il permet l'intégration des essais expérimentaux et des données environnementales. Mais il permet aussi de vérifier la qualité des données en établissant des liens avec d'autres systèmes qui référencent les ressources génétiques (ex : les données passeport). De plus, il dispose de dictionnaires de variables liés aux ontologies agronomiques notamment pour les espèces de maïs, tomates et vigne (C. Pommier, 2009).

Cependant, ce système ne permet pas le suivi des interventions techniques réalisées sur un essai ni le suivi des échantillons. Par ailleurs, le système ne permet pas la caractérisation des organes ou tout autre entité se trouvant à une échelle inférieure à celle d'une unité expérimentale. Ephesis ne permet pas non plus le suivi des informations de généalogies liées aux génotypes. En plus de ces aspects fonctionnels, il existe des aspects techniques qui posent problème. Ephesis est né d'un besoin institutionnel de l'INRA, il est par nature difficilement transférable à un autre institut. De plus, il est toujours en cours de développement ce qui engendre des difficultés pour travailler correctement.

LE BREEDING MANAGEMENT SYSTEM DE L'IBP

L'IBP (The Integrated Breeding Platform), a pour objectif d'accélérer la création et la distribution de variétés de plantes améliorées pour répondre à la demande alimentaire croissante dans le monde.

Pour ce faire, la plateforme fournit des outils et des services professionnels pour la conduite d'activités de phénotypage et de génotypage, dont le BMS (Breeding Management System). Le BMS est un ensemble d'outils logiciels qui intègre une base de données permettant de gérer les informations de pedigree, de caractérisation phénotypique et d'évaluation de matériel végétal. De plus, il inclut des dictionnaires de caractères « ontologiques » pour 10 espèces dont le sorgho, un gestionnaire de ressource génétique et d'essais ainsi qu'un outil permettant le suivi des croisements.

Toutefois, ce système ne permet pas l'intégration des données phénotypiques historiques, ce qui est un point important pour des projets tels que BFF qui possèdent déjà des jeux de données complets répartis sur plusieurs années.

De plus, L'architecture du BMS ne permet pas de travailler sur plusieurs cultures avec une seule instance du système. Chaque instance du système est dédiée à une seule espèce. A l'instar d'Ephesis, le BMS ne permet pas la gestion des observations à l'échelle des organes et c'est un système distribué gratuitement mais pas *open-source* on ne peut donc pas le faire évoluer librement.

PHENOME NETWORKS ET AGROBASE

D'autres systèmes tel que Agrobases ou Phenome Networks présente des fonctionnalités intéressantes, ce sont des systèmes d'informations orientés vers la sélection variétale et qui intègrent tout deux des outils pour l'identification des marqueurs génétiques affectant les traits étudiés. Néanmoins, ces systèmes sont complètement opaques de par leur nature commerciale ce qui empêche toute modification visant à améliorer ces systèmes sans intervention de leurs propriétaires. De plus, ils ont un coût financier notable.

CONCLUSION

En conclusion, chaque système d'informations développé pour la gestion des essais agronomiques propose des fonctionnalités intéressantes. En revanche, la majorité de ses solutions nous confronte aux difficultés de maintenance et d'évolution de ces systèmes, critères importants pour leur pérennité au sein du CIRAD ; ainsi qu'au manque de fonctionnalité (ex : gestion des organes, des données historiques, gestion de plusieurs espèces... etc.).

DAPHNE

Il existe des outils qui répondent aux besoins de stockage et d'exploitation des données génétiques et génomiques. Cependant, la compréhension du phénotype nécessite l'intégration de ces données mais aussi, des données expérimentales.

C'est pourquoi, Les équipes du projet BFF ont décidé de développer un système générique pour la gestion des données phénotypiques et environnementales. C'est dans ce but que j'ai été recruté en tant que stagiaire. BFF est un cas d'étude pour le développement de ce système, l'idée étant de le déployer plus largement. Cet outil doit servir à d'autres projets agronomiques que BFF et à l'étude d'autres espèces que sorgho et miscanthus.

LE CAHIER DES CHARGES

Les objectifs de ce stage sont doubles :

- Construire une base de données permettant de centraliser, stocker et sécuriser les données actuelles produites dans le cadre du projet BFF et à venir, issues des différents domaines suivants : agronomie, écophysiologie, météorologie, génétique, sélection, histologie, biochimie.
- Développer une application web permettant l'alimentation et l'interrogation de cette base de données par les différents partenaires publiques et privés.

LES CARACTERISTIQUES DU SYSTEME D'INFORMATIONS

Ce système de gestion des données doit intégrer :

- La description des dispositifs expérimentaux.
- La gestion des itinéraires techniques
- La gestion des données météorologiques et sol
- La gestion des échantillons produits
- La description des ressources génétiques
- La gestion des résultats d'analyses
- La gestion des observations à toutes les échelles

Il doit aussi prévoir l'interopérabilité avec les ontologies existantes (PLANT ONTOLOGY, CROP ONTOLOGY, AGRONOMY ONTOLOGY... etc.) et les collections de ressources génétiques du CRB (Centre de Ressource Biologique). De plus, le système doit aussi permettre l'intégration des données correspondant à plusieurs espèces dans un même essai pour pouvoir décrire des dispositifs de culture en association.

En relation avec les attentes du système d'information, J'ai proposé le nom **DAPHNE** (**DA**tabase **PH**enotype **plaNt** **intE**gration) qui a été retenu.

MATÉRIELS ET MÉTHODES

VOLUMETRIE ET COMPLEXITE DES DONNEES

Les données produites par un centre de recherche agronomique sont de nature très hétérogène. Rien que pour le projet BFF qui se cantonne à l'étude de 2 espèces, c'est l'analyse d'environ 3000 génotypes en semis, répartis sur une trentaine d'essais et sur 10 lieux d'expérimentations depuis 2012 (Figure 2).

Chacun des dispositifs expérimentaux mis en place peut être plus ou moins complexe et définit sur plusieurs niveaux (blocs, parcelles, places, placettes, etc.). Sur chacune des unités expérimentales que décrit le dispositif, s'applique une ou plusieurs modalités (Figure 3). Chaque essai est mené en suivant un itinéraire technique qui comprend toutes les interventions réalisées sur chacune des unités du dispositif (fertilisation, travail du sol, irrigation, etc.).

Ajoutés à cela, les essais agronomiques recensent des données liées aux caractéristiques des sols qu'ils soient géo-localisés ou non (profondeur, type de sol, humidité) et des données liées aux conditions climatiques (température, humidité, vent, rayonnement) récoltées toutes les minutes par des capteurs météorologiques. Ils récoltent aussi des données au champ sur une cinquantaine de caractères liés aux phénotypes des plantes (hauteurs, diamètres, nombre de feuilles, nombre d'entrenœuds, etc.) ou à leur phénotype étendu (ex : taux d'envahissement par les mauvaises herbes) (Figure 4).

De plus, ces études agronomiques génèrent des quantités importantes d'échantillons récoltés à toutes les échelles et à différents stades. Ces échantillons sont transférés à de nombreux laboratoires pour les analyser sur une trentaine de caractères histologiques, physiologiques et biochimiques.

	A	B	C	D	E	F	G	H	I	
	Accession_name_BFF	Accession_name_BFF_Synonyms	Genotype_code_BFF	Accession_BFF_code_database_Old	Genotype_Type	Female_Parent_Accession_name_BFF	Male_Parent_Accession_name_BFF	Female_Parent_Genotype_Code_BFF	Male_Parent_Genotype_Code_BFF	A/B_R
1										
2	RE1xAE1	AE1_X_RE1	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
3	RE1xAE1	AE1XRE1	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
4	RE1xAE1	Biomass 140	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
5	RE1xAE1	Biomass140	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
6	RE1xAE1	Biomasse140	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
7	RE1xAE1	BIOMASSE140=HERCULES	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
8	RE1xAE1	H1	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
9	RE1xAE1	Hercules	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
10	RE1xAE1	RE1xAE1	G00001	G00001	F1_Hybrid	AE1	RE1	G00139	G00158	Not_re
11	RE1xAE4	AE4_X_RE1	G00002	NA	F1_Hybrid	AE4	RE1	G00141	G00158	Not_re
12	RE1xAE4	RE1xAE4	G00002	G00002	F1_Hybrid	AE4	RE1	G00141	G00158	Not_re
13	RE1xAE4	W234F506	G00002	NA	F1_Hybrid	AE4	RE1	G00141	G00158	Not_re
14	RE1xAR3	AR3_(BF003)_X_RE1	G00003	NA	F1_Hybrid	AR3	RE1	G00144	G00158	Not_re
15	RE1xAR3	RE1xAR3	G00003	G00003	F1_Hybrid	AR3	RE1	G00144	G00158	Not_re
16	RE1xAR3	W234F533	G00003	NA	F1_Hybrid	AR3	RE1	G00144	G00158	Not_re
17	RE1xAR4	AR4_X_RE1	G00004	NA	F1_Hybrid	AR4	RE1	G00145	G00158	Not_re
18	RE1xAR4	RE1xAR4	G00004	G00004	F1_Hybrid	AR4	RE1	G00145	G00158	Not_re
19	RE1xAR4	W234F535	G00004	NA	F1_Hybrid	AR4	RE1	G00145	G00158	Not_re
20	IS19453	G13_BFPN_020	G00005	G00005	Pure_Line	IS19453	IS19453	G00005	G00005	
21	IS19453	IS 19453	G00005	G00005	Pure_Line	IS19453	IS19453	G00005	G00005	
22	IS19453	IS19453	G00005	G00005	Pure_Line	IS19453	IS19453	G00005	G00005	
23	IS19453	IS19453_X_IS19453	G00005	NA	Pure_Line	IS19453	IS19453	G00005	G00005	
24	IS20351	G13_BFPN_037	G00006	G00006	Pure_Line	IS20351	IS20351	G00006	G00006	

Figure 2 Exemple de fichier contenant des données liées aux accessions des génotypes du projet BFF pour l'espèce sorgho.

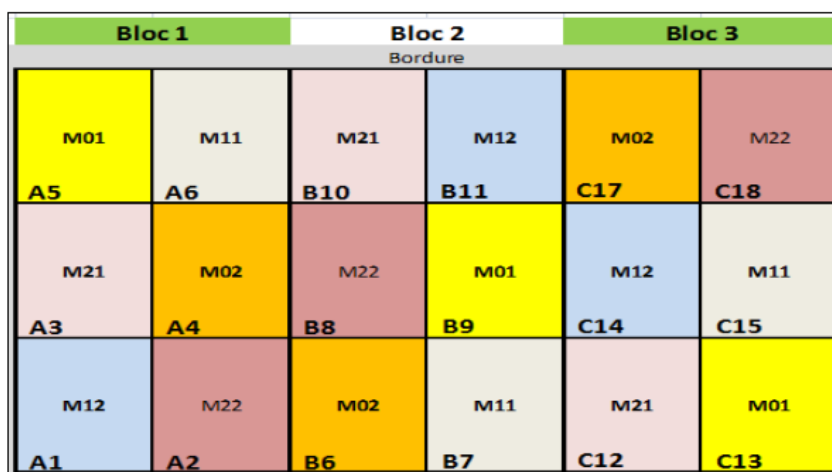


Figure 3 Exemple de schéma d'un dispositif expérimental.
An, Bn, Cn : parcelles ; Mn : modalités ; Blocn : répétition

APPROCHE DE DÉVELOPPEMENT

Qu'il s'agisse de modélisation ou de conception de l'interface web et de la base de données, nous avons mis en pratique une organisation agile de développement très similaire à la philosophie de la méthode SCRUM (Ken Schwaber, 2001). Au cours de réunions d'équipe hebdomadaires, nous avons défini des tâches à réaliser à la suite d'un échange d'idées avec les encadrants du stage et des experts des différents domaines abordés par le projet. Ces intervenants ont souvent éclairé nos débats afin de proposer les meilleures solutions pour répondre aux attentes des utilisateurs notamment sur la gestion des données d'accessions ou encore les notions de vocabulaire contrôlés en lien avec les ontologies (phénotypique, agronomique, génétique)

LA BASE DE DONNÉES

A l'instar d'ECOFI, il a été décidé d'implémenter la base de données sous PostgreSQL. C'est un système de gestion de base de données relationnelle-objet (SGBDRO). Il est puissant et *open-source*. C'est aussi l'un des SGBD les plus stables sur le marché. (Voir en annexe 1, la liste des outils utilisés pour le développement de la base de données)

PRÉSENTATION DES PACKAGES ET CAS D'UTILISATIONS

La première étape pour concevoir un système d'information est la modélisation de la base de données. Pour ce faire il est important de définir ce que le système devra ou non gérer comme données. On a défini des sous-systèmes (packages) qui permettent de regrouper la structure de la base de données par champ thématiques (Figure 5). Chacun de ces packages répond à un ou plusieurs cas d'utilisations. Ils ont servi de guide pour construire le schéma relationnel de la base.

Gestion des utilisateurs et des accès

Une composante importante du système d'information est l'authentification de l'utilisateur, pour pouvoir lui donner un accès et lui attribuer des droits en lecture, écriture ou modification des données contenues dans la base. Pour cela il doit s'authentifier à l'aide d'un nom d'utilisateur et d'un mot de passe. Il doit aussi pouvoir partager et contrôler la visibilité de ses données. Il peut créer un jeu de données et décider de qui peut lire, ajouter ou modifier les données.

	A	B	C	D	E	F	G	H	I	J	K	L	M
					Nbre de Feuilles ligulées. Plantes Biomasse	Nbre de Feuilles vivantes. Plantes Biomasse	Longeur N- 4 (cm)	Largeur N-4 (cm)	Hauteur dernière ligule (cm). Plantes Biomasse	Hauteur base péduncule (cm)	diamètre à la base (mm). Plantes Biomasse	Nombre d'entre-nœud allongés . Plantes Biomasse	Remarque
1	Plot	Harvest_date	Plant_Number	Trait_Status									
2	P1013	13/08/2015	1	measured	18	8	268	277	76	7,5	16,	10	
3	P1013	13/08/2015	2	measured	18	7	240	252	71	6,5	15	10	
4	P1013	13/08/2015	3	measured	18	8	265	265	73	7,5	16,6	9	
5	P1013	13/08/2015	4	measured	18	8	262	268	70	6,7	14,9	10	talles axillaires
6	P1028	13/08/2015	1	measured	17	8	195	240	NA	NA	19,2	10	talles axillaires
7	P1028	13/08/2015	2	measured	16	7	218	221	76	6,5	17,5	9	talles axillaires
8	P1028	13/08/2015	3	measured	17	7	217	240	NA	NA	18,33	10	talles axillaires
9	P1028	13/08/2015	4	measured	17	7	254	268	78	6,6	14,14	10	talles axillaires
10	P1098	13/08/2015	1	measured	17	7	243	254	87	9	19,9	10	
11	P1098	13/08/2015	2	measured	19	8	239	252	60	8	19,75	10	
12	P1098	13/08/2015	3	measured	18	7	244	257	88	8,7	20,2	10	
13	P1098	13/08/2015	4	measured	18	7	265	282	85	7,4	17,7	10	
14	P1100	13/08/2015	1	measured	16	7	246	262,5	80	6,5	18	9	
15	P1100	13/08/2015	2	measured	16	6	227	245	85	7	14,2	8	
16	P1100	13/08/2015	3	measured	16	6	226	246	85	6,9	15,9	8	
17	P1100	13/08/2015	4	measured	15	7	216	235	78	6,2	15,4	9	
18	P1107	13/08/2015	1	measured	16	6	257	272	70	7,5	14,3	9	
19	P1107	13/08/2015	2	measured	16	6	255	270	72	7,5	15	9	
20	P1107	13/08/2015	3	measured	17	7	258	280	51	6,4	14,3	10	
21	P1107	13/08/2015	4	measured	16	7	246	260	59	8	15,3	9	
22	P1108	13/08/2015	1	measured	16	7	236	244	78	6,4	15,35	9	
23	P1108	13/08/2015	2	measured	16	7	242	253,5	77	6,5	16	9	
24	P1108	13/08/2015	3	measured	16	8	238	250	73	6,5	14,9	9	
25	P1108	13/08/2015	4	measured	16	7	235	256	78	6,1	15,09	9	

Figure 4 Exemple de fichier contenant des données morphologiques

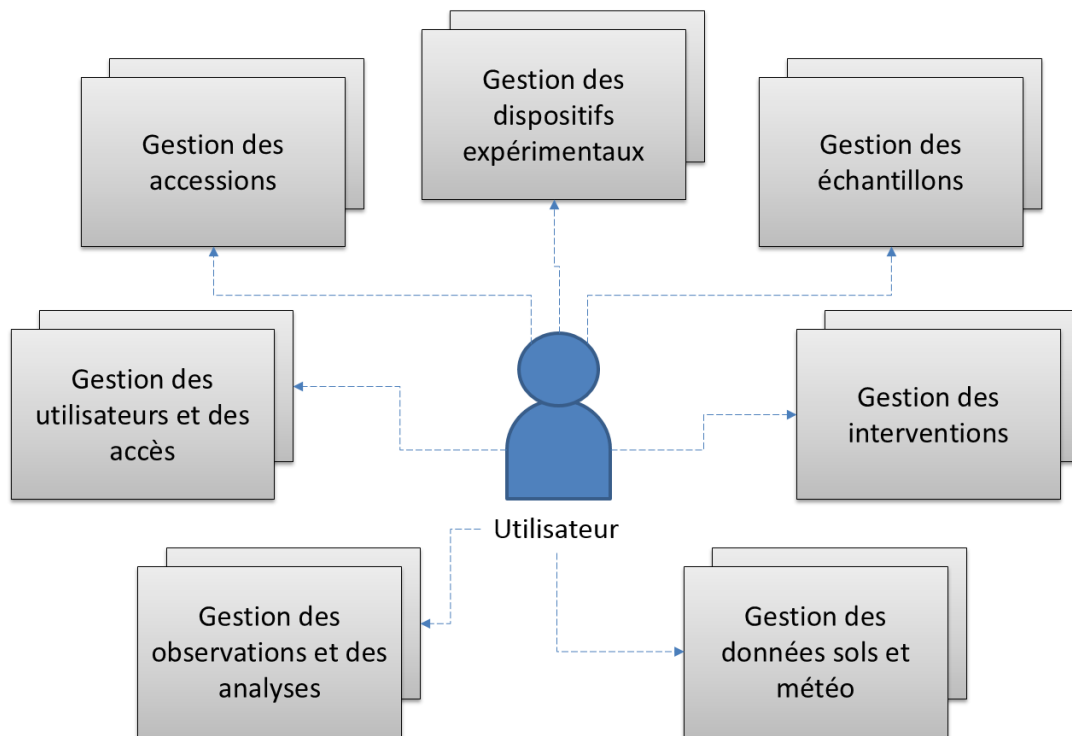


Figure 5 Représentation des packages principaux du système d'information

Gestion des dispositifs expérimentaux

L'utilisateur doit pouvoir préciser le dispositif expérimental d'un essai mis en place au champ. Pour cela il doit définir le projet dans lequel cet essai a lieu, la localisation de l'essai et décrire le dispositif ainsi que les modalités (traitements) affectées à chacune des unités expérimentales (parcelles, placettes, etc.).

Gestion des accessions

L'utilisateur doit pouvoir ajouter les géotypes qui font partie des modalités d'un ou plusieurs dispositifs. Pour ce faire, il doit renseigner des accessions qui seront stockées dans la base de données. Une accession est définie par un taxon (le plus souvent une variété ou une espèce) produit sur un lieu donné et à une date donnée.

Gestion des échantillons

L'utilisateur doit pouvoir suivre les échantillons issus d'une parcelle, d'une placette ou d'une plante individuelle d'un essai tout le long de leur durée de vie. Il devra pouvoir avoir accès à tout moment au lieu et au partenaire qui détient l'échantillon et tracer les divisions et toutes les opérations qu'aura subi l'échantillon.

Gestion des observations et des analyses

L'utilisateur doit pouvoir stocker tous les résultats d'analyses (biochimique, génomique, etc.) et toutes les observations possibles réalisées à tous les niveaux hiérarchiques de l'essai (plante, groupe de plantes, placette, parcelle, etc.) Il est important qu'il puisse déclarer ses propres variables si elles n'existent pas déjà dans la base de données.

Gestion des données météorologiques et sol

L'utilisateur doit pouvoir stocker les données météorologiques récoltées par les différents types de stations météo. A la mise en place de l'essai, il doit pouvoir cartographier le sol en définissant le type de sol, sa profondeur et les différentes couches qui le constituent et réaliser le suivi des couches dans le temps à partir des observations.

L'utilisateur doit pouvoir suivre l'itinéraire technique d'une unité expérimentale. Pour cela il doit pouvoir ajouter n'importe quel type d'intervention (ex : traitement phytosanitaire, épandage d'engrais, récolte) réalisée sur une ou plusieurs unités expérimentales et spécifier la date de l'intervention et sa durée. Il doit aussi pouvoir tracer le dosage qui a été appliqué si l'intervention consiste en l'application d'un produit (ex : La diffusion d'un traitement chimique contre un parasite sur une parcelle donnée).

LE FRAMEWORK

L'un des objectifs principaux de l'application est l'intégration des données via des formulaires en ligne. Pour ce faire il a été décidé en début de projet, d'utiliser le Framework PHP Phalcon. Les Framework offrent une méthode de travail structurée permettant de maintenir facilement un projet.

Phalcon est un Framework PHP 5 open source performant. Il est écrit en C, et est disponible en tant qu'extension PHP. L'avantage de Phalcon est qu'il intègre un module de gestion des utilisateurs et des droits accès. Cependant, le recul qu'a apporté le développement de l'interface d'ECOFI avec ce Framework a montré son manque de souplesse et sa limitation vis-à-vis d'autres Frameworks sur le marché. Avec les contraintes de temps qu'imposent un stage de fin d'étude et la volonté de simplifier la maintenance des outils informatiques du centre de recherche, il était préférable de se tourner vers un Framework plus adapté au type de développement envisagé et plus simple d'apprentissage. C'est pour ces raisons et en nous fondant sur comparatif réalisé au travers d'entretiens et de retours de praticiens sur le développement d'application web, que nous avons finalement choisi le Framework CodeIgniter.

CodeIgniter 3 est un Framework PHP qui s'exécute du côté du serveur de l'application (back-end). Il est constitué d'un ensemble de ressources pour permettre un développement rapide, sécurisé et optimisé. Il possède une documentation claire et complète. C'est l'un des Framework PHP les plus légers actuellement (espace disque requis : 2 MB). De plus il offre une simplicité d'utilisation peu commune face à des concurrents tel que Symphony, Phalcon ou encore CakePHP. Il utilise l'architecture logicielle MVC qui permet de faciliter grandement le développement et la maintenance des applications web (Reenskaug, 1979).

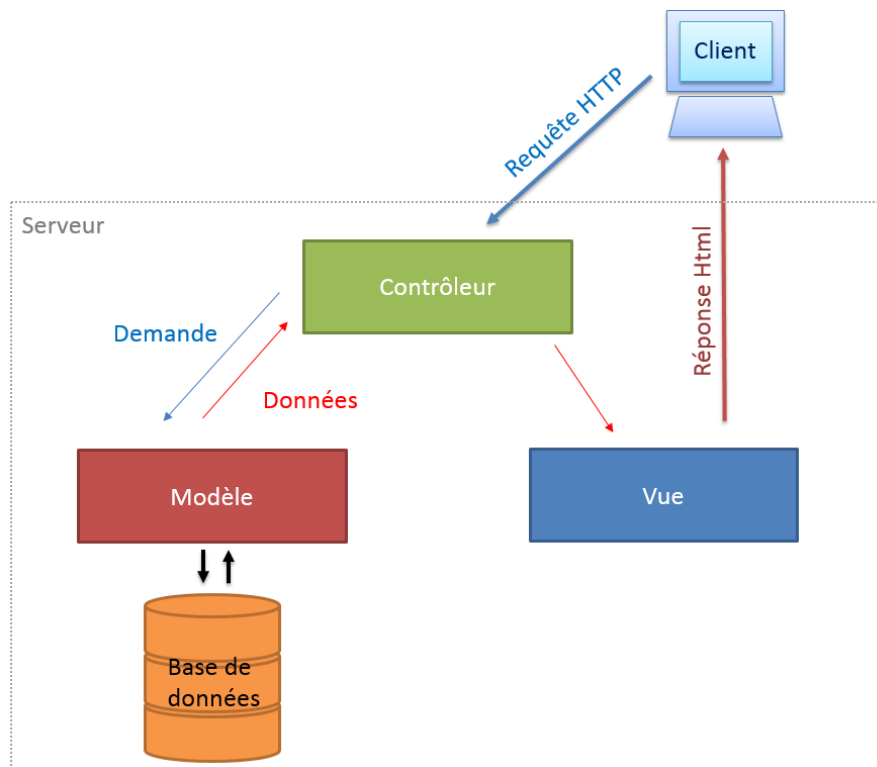


Figure 6 Schéma de l'architecture MVC

```

Welcome.php  Member.php

1  <?php
2  defined('BASEPATH') OR exit('No direct script access allowed');
3  /**
4   * Contrôleur principal de L'application web DAPHNE
5   * @author Medhi Boulnemour <boulnemour.medhi@live.fr>
6   */
7  class Welcome extends MY_Controller {
8      /**
9       * Constructeur du contrôleur
10      * Il est exécuté avant d'appeler les pages de du contrôleur 'Welcome'
11      */
12      public function __construct() {
13          parent::__construct();
14          $this->load->model('users_model');
15          $this->load->model('Group_model');
16          $this->load->library('form_validation');
17          $this->load->helper('html');
18
19          //Vérifie si l'utilisateur est connecté
20          if($this->session->userdata('connected') == TRUE)
21          {
22              //Enregistre les groupes et les projets associés à l'utilisateur
23              $this->load->model('Group_model');
24              $this->user_groups = $this->Group_model->get_user_groups($this->session->u
25          }
26      }
27      /**
28      * Index Page for this controller.
29      */
30      public function index() {
31          $this->view();
32      }

```

Figure 7 Le contrôleur de la page d'accueil de l'application – Capture d'écran de l'éditeur de code Atom

MODÈLE-VUE-CONTROLEUR

Le patron d'architecture logicielle MVC, est un paradigme de programmation qui sépare les problématiques et les fonctions nécessaires pour y répondre en trois parties (Figure 6) (Reenskaug, 1979).

LE CONTROLEUR

Le contrôleur s'affaire à la gestion des événements. C'est la classe qui est appelée en premier lorsque l'utilisateur émet une requête http. Le contrôleur appelle les bibliothèques, les vues, les modèles et réalise les vérifications des formulaires. Un contrôleur dans CodeIgniter est une classe qui en hérite par défaut de la classe CI_Controller (Figure 7). Toutes les méthodes publiques définies dans le contrôleur sont autant de pages accessibles via l'URL.

Exemple : <http://daphne.CIRAD.fr/membre/profil/Paul>

Racine de l'application	DAPHNE.CIRAD.fr
Contrôleur	Membre
Méthode	Profil
Paramètre	Paul

LA VUE

La vue est l'interface utilisateur à proprement parler. Cette partie est responsable de tout ce qui sera affiché à l'écran. Afin de faciliter le développement et gérer le design de l'application, on a choisi le Framework Bootstrap.

Bootstrap est un Framework côté client (front-end) développé avec CSS3, HTML5 et JavaScript. Il contient une bibliothèque sous licence Apache et totalement open source.

LE MODÈLE

Un modèle est une classe représentant une partie des données d'une application. C'est lui qui fait le lien entre l'application web et la base de données. Les Modèles héritent de la classe MY_Model qui propose des fonctionnalités de type CRUD (Create, Read, Update, Delete) et permet de faciliter l'utilisation de la base de données. Ces méthodes permettent de diminuer le temps passé à la création des requêtes.

SYTÈME DE GESTION DE VERSIONS : GIT

Le suivi de développement de l'application web s'est effectué via le Gitlab du CIRAD, ce qui nous a permis de gérer le stockage, l'archivage et les modifications des sources de l'application. Gitlab est un logiciel libre de gestion des dépôts Git. Il gère aussi les utilisateurs et leurs droits d'accès aux dépôts. Ce qui permet de collaborer à plusieurs sur le développement d'un projet ainsi que de partager ce projet publiquement ou uniquement à certains utilisateurs. Pour des raisons de sécurité propre au développement d'une application web liée à une base de données, ce dépôt Git est privé et partagé uniquement entre les membres du projet BFF en charge de son développement.

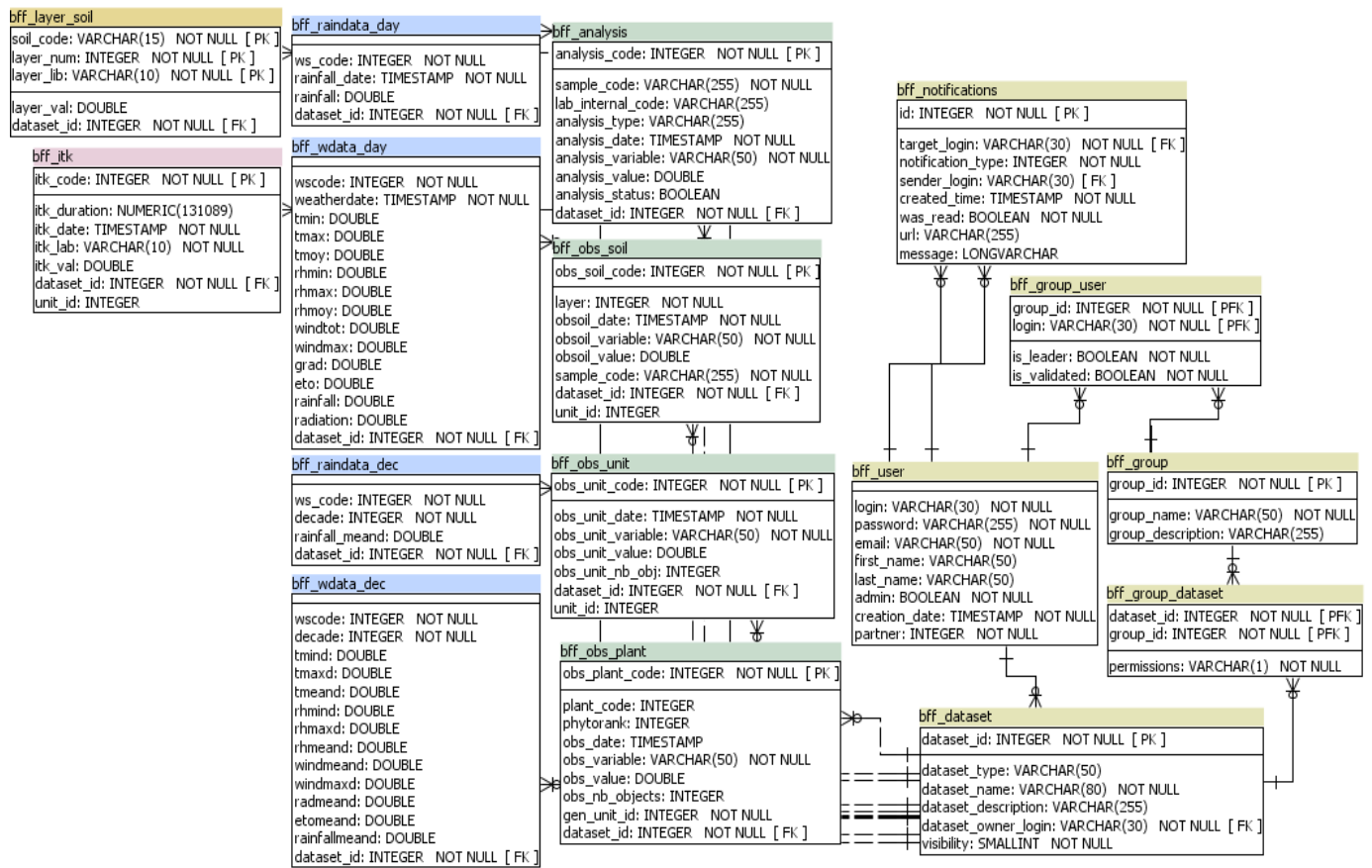


Figure 8 Schéma UML du module de gestion des utilisateurs et des accès aux jeux de données (en vert olive, à droite) ; Les tables pouvant appartenir à un jeu de données (en bleu : les données météorologiques ; en rose : les données des itinéraires techniques ; en marron : les données sol ; en vert : les données d'observation et d'analyse)

RÉSULTATS ET DISCUSSION

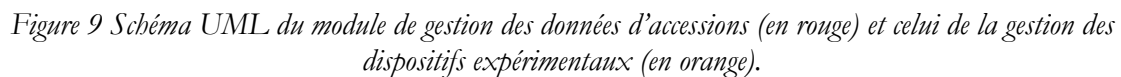
IMPLÉMENTATION DE LA BASE DE DONNÉES

La structure de la base de données implémentée [Annexe : 3] se divise en 7 parties :

1. La partie gestion des utilisateurs, accès et administration. Cette partie de la base de données est construite autour de 3 tables principales :

- La table « bff_user » stocke les informations relatives aux utilisateurs (ex : mots de passe, adresse email, etc.) et indique si ces derniers sont administrateurs du système. Les mots de passe sont hachés avant l'enregistrement dans la base de données par l'application web via une fonction PHP qui utilise par défaut l'algorithme de hachage Blowfish.
- La table « bff_group » permet de définir des groupes d'utilisateurs. Des droits en lecture ou en écriture sur les jeux de données seront attribués à chaque groupe. En fonction de ces droits Les utilisateurs pourront lire, ajouter ou modifier des données auxquelles ils auront accès.
- La table « bff_dataset » permet de définir un jeu de données et d'en contrôler l'accès. On peut définir l'accès à un jeu de données soit de manière standard en choisissant de le rendre privé ou publique. Soit de manière plus précise en définissant quelle sont les groupes d'utilisateurs qui auront accès à un jeu de données et de quelle manière (lecture ou écriture)

Pour développer ce module, je me suis beaucoup intéressé à la façon dont les accès aux fichiers étaient gérés sur les système Unix, c'est d'ailleurs de là que m'est venue l'idée d'utiliser des groupes d'utilisateurs dans DAPHNE. Avec cette notion de groupe transverse aux notions de partenaires et de projets, les utilisateurs peuvent décider précisément qui aura accès à leurs données (Figure 8).



2. La partie dispositif expérimental décrit le design d'un essai mis en place dans un champ (Figure 3). La table principale de cette partie, est la table « exp_unit ». Une relation récursive permet de définir autant de niveaux hiérarchiques que l'on veut pour décrire un dispositif. Chaque entrée de cette table peut faire référence à une autre entrée de cette même table via une clé étrangère. Un même dispositif expérimental peut être utilisé dans plusieurs projets de recherches (Figure 9).
3. La partie gestion des accessions stocke les informations liées aux génotypes qui seront mis au champ dans les différents essais définis dans la partie "dispositif expérimental" de la base de données. La table « bff_accession » permet de définir un génotype produit à un lieu donné et à une date donnée (Figure 9). C'est de cette manière que sont identifiés les lots de semences et les génotypes par les autres systèmes d'informations du CIRAD. En épousant cette structure de données pour définir une accession cela permet d'envisager l'interopérabilité avec ces systèmes d'informations qui gère le stockage des données génomiques.
4. La partie gestion des données sol et météo stocke les données environnementales issues des différents capteurs météorologiques ou pédologiques utilisés dans les essais au champ (Figure 10).
5. La partie gestion des observations et analyses permet d'enregistrer les données liées aux observations et les analyses réalisées au champ à tous les niveaux (plante, placette, parcelles, etc.). Afin que les données recueillies puissent être comparables entre elles, il est important de définir leurs natures. Pour cela il existe de nombreux projets visant à bâtir des ontologies en agronomie. Cependant ces ontologies sont souvent incomplètes et ne permettent pas à elles seules de définir un dictionnaire de variables exhaustif de toutes les observations réalisables sur une ou plusieurs espèces. C'est pourquoi nous avons construit un système de tables directement empruntée à la base de données ECOFI et permettant de définir une variable d'observation à n'importe quels niveaux de l'essai si elle n'est pas déjà définie dans la base de données. Cela permet d'enrichir le dictionnaire de variables de la base de données. Cette variable est définie par le trait observé, la méthode de mesure, calculée ou d'observation, l'unité de mesure et l'entité observée (Figure 11). De plus, il faut prendre en considération que les ontologies existantes continuent d'évoluer.

Ainsi, le système de tables qui gère les données du vocabulaire contrôlé contient une table « bff_ontology » qui référence les ontologies existantes lié aux variables d'observations. De cette manière on peut envisager l'interopérabilité avec d'autres systèmes d'informations qui recense des variables d'observations phénotypiques.

6. La partie gestion des échantillons permet la traçabilité des échantillons biologiques. J'ai construit les 3 tables suivantes : « bff_sample » qui permet de décrire un échantillon ; « bff_sample_stage » qui permet de définir les différents stades de prélèvements ; « bff_transfer » est la table qui référence les échanges d'échantillons entre les différents laboratoires et permet le suivi des échantillons (Figure 12).
7. La partie gestion des interventions est la partie du système d'information qui trace l'ensemble des interventions effectuées au champ sur une unité expérimental donnée (ex : La pulvérisation d'un traitement chimique sur une parcelle). Elle est constituée d'une table unique « bff_itk » en relation avec la table qui décrit les unités expérimentales « bff_exp_unit » (Figure 12). Ainsi, on peut décrire une intervention du plus haut niveau comme à la plus petite échelle d'un dispositif.

IMPLÉMENTATION DE L'APPLICATION WEB

La structure de l'application est construite selon l'architecture à trois couches, comme présentée sur la Figure 13. [Annexe : 2]

FONCTIONNALITES

L'application est actuellement en ligne sur le serveur Alberes. Pour permettre aux utilisateurs d'accéder aux données de DAPHNE ou de partager de nouvelles données, ils doivent passer par une étape d'authentification via l'interface de connexion (Figure 15).

Connexion

Pour se connecter à l'application, l'utilisateur fournit un nom d'utilisateur et un mot de passe valides. Ces données sont stockées dans la table « bff_user » après avoir renseigné le formulaire d'inscription. Les mots de passe sont hachés avant d'être stockés dans la base de données en utilisant l'algorithme de hachage « Blowfish » (Schneier, 1993) via la fonction PHP « password_hash ».

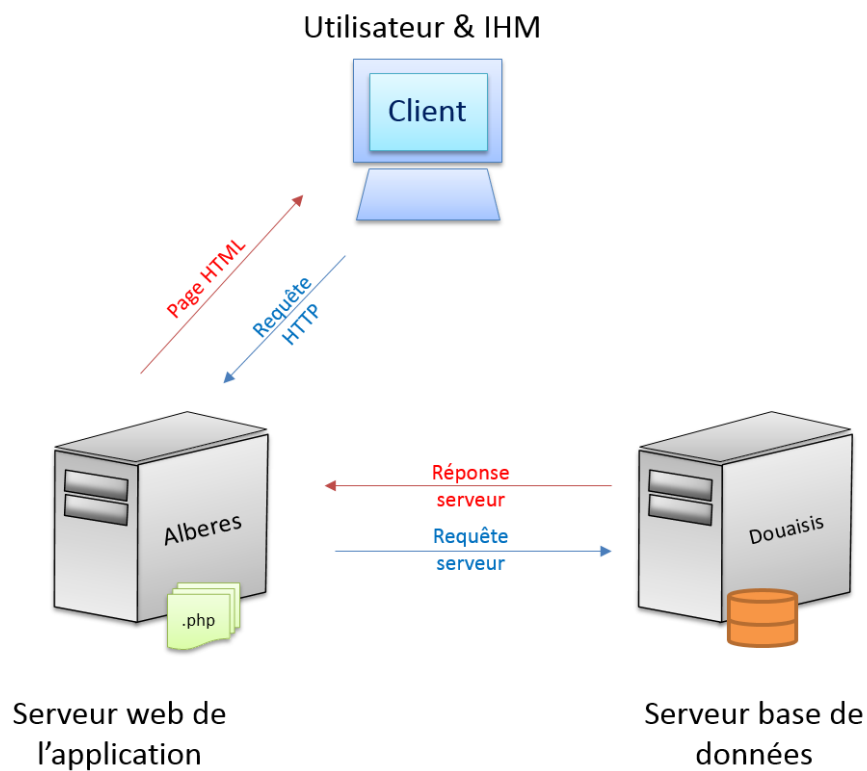


Figure 13 L'architecture à trois niveaux du système d'information

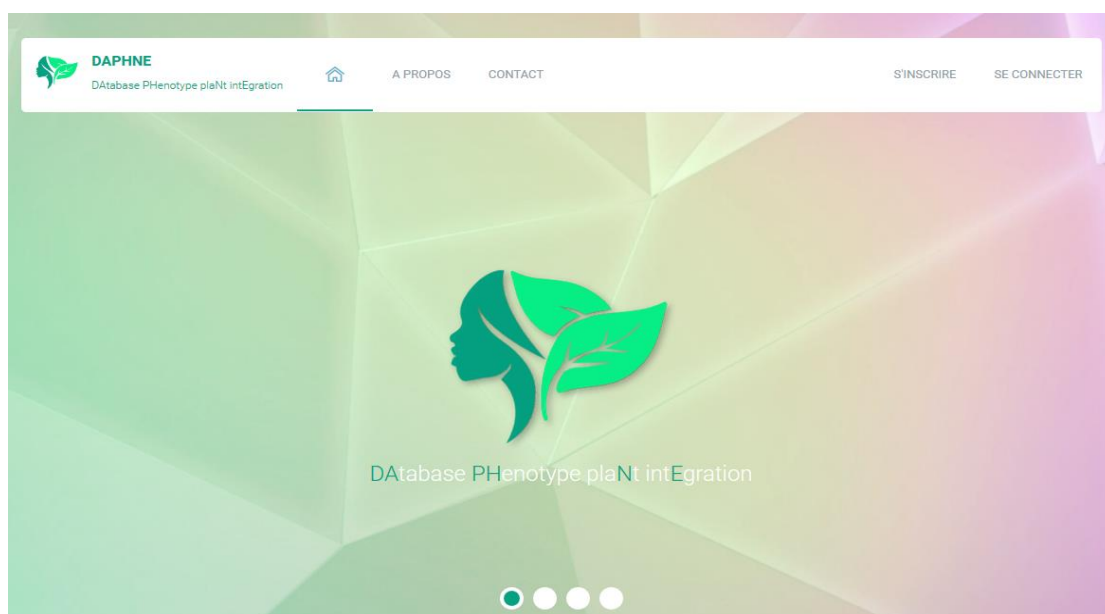


Figure 14 Page d'accueil de l'application web DAPHNE

A l'issue de cette étape d'authentification, l'utilisateur a accès à un espace membre (Figure 16). Cette espace permet l'accès à différentes sections de l'application :

- L'accès à la gestion des données. L'utilisateur peut créer un jeu de données et consulter les données aux quelles lui ou les groupes dont il fait partie ont accès.
- L'espace d'administration de DAPHNE. Cette espace est uniquement réservée aux administrateurs de la base de données. Il s'agit d'un espace qui permet la visualisation, l'ajout et la modification des données de DAPHNE. Il permet aussi la validation des projets et propose une interface simplifiée pour gérer les utilisateurs de la base de données.
- L'accès à la gestion/création de groupe. L'utilisateur peut créer des groupes afin de partager ces données. Pour cela il peut définir des droits en lecture ou écriture à autant de groupes qu'il le souhaite sur un ou plusieurs ses jeux de données.
- L'accès à la gestion/création de projet. Les données issues des essais sont rattachées à des projets. Les utilisateurs doivent donc renseigner le projet dans lequel l'essai est effectué avant de pouvoir alimenter la base de données.

Importation de données

L'intégration des données s'effectue via l'interface de saisie et d'importation des données. Cette page web est un formulaire dynamique qui se renseigne en 3 étapes :

1. La première étape consiste à choisir le formulaire qui correspond aux données qu'on souhaite stocker dans DAPHNE (Figure 17). Il existe 2 types de formulaires. Certains sont uniquement publics, comme le design du dispositif expérimental. D'autres doivent être reliés à un jeu de données et constitue un capital qui peut être défini comme publique ou privé. Ces jeux de données peuvent aussi être partagés uniquement avec certains groupes d'utilisateurs que le propriétaire des données aura défini au préalable.
2. Après avoir choisi un formulaire, l'interface fait apparaître les champs correspondants. Certains formulaires proposent d'importer les données au format CSV (comma separated values) et pré-visualiser le fichier avant la soumission.

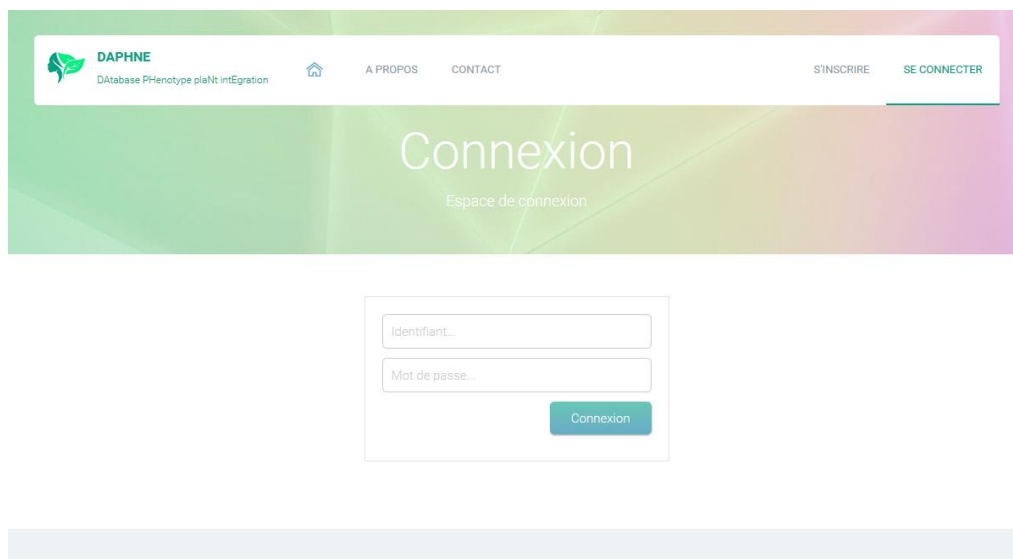


Figure 15 Interface de connexion de DAPHNE

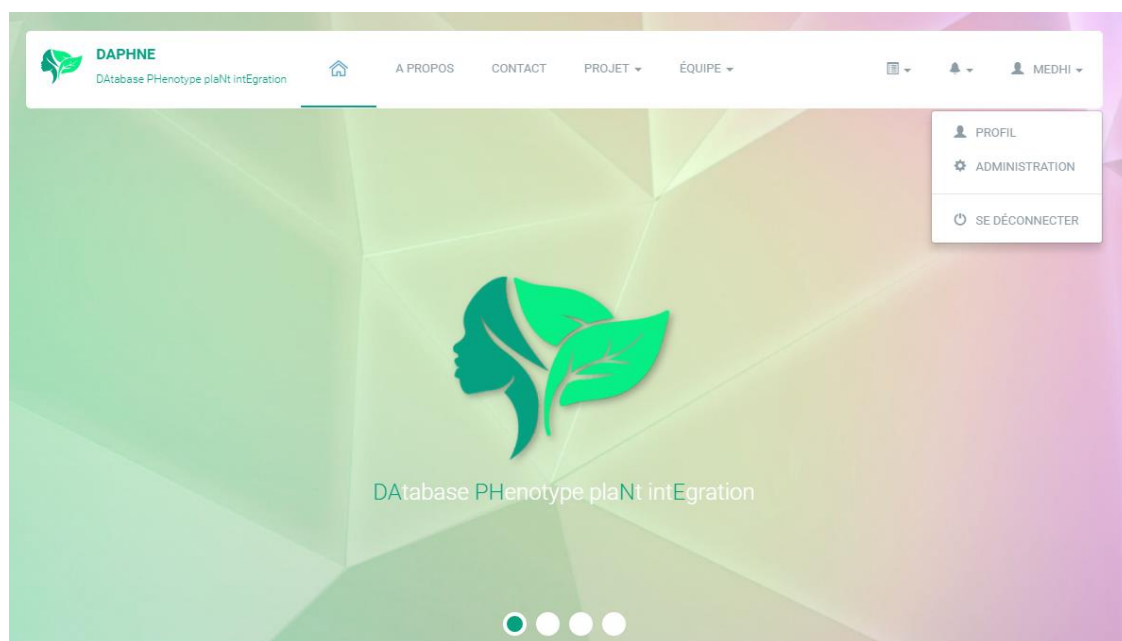




Figure 16 Espace membre matérialisé par de nouveaux liens disponibles dans la barre de navigation. L'icône  permet l'accès à la page de saisie et d'importation de données ainsi qu'aux données de l'utilisateur. L'icône  permet l'accès aux notifications. Ces notifications informent l'utilisateur des différents évènements le concernant (ex : invitation à faire partie d'un groupe)

3. Pour les formulaires de données à accès contrôlé, il faut définir à quel jeu de données appartiennent les informations renseignées. Pour cela, l'interface propose de sélectionner un jeu de données existant ou bien d'en créer un nouveau et définir la visibilité du jeu de données (publique, privée ou partagée avec des groupes spécifiques).

DIFFICULTÉS RENCONTRÉES

LES OBJECTIFS

Avant mon arrivée dans l'équipe, mes objectifs de stage devaient se cantonner à la réalisation d'une interface web pour l'alimentation et l'interrogation de DAPHNE, puis de développer un module permettant le suivi des échantillons biologiques conforme à l'outil prés-existant développé avec Microsoft® Excel. Cependant, forcé de constater que la construction de la base de données n'était pas achevée à mon arrivée au CIRAD, nous avons décidé de réévaluer les attentes du stage. J'ai donc participé activement au développement de la base de données sur tous les aspects, et pas uniquement à la gestion des échantillons ; ainsi qu'au développement de l'application web en sachant qu'il serait difficile de terminer dans le délai imparti.

L'AGILITÉ

Nous avons décidé de concevoir parallèlement la base de données et l'application web en suivant une méthode agile. C'est une approche intéressante de développement mais qui a pu poser parfois des difficultés au niveau de la programmation de l'application. En effet lorsqu'on effectue des changements importants dans la base de données, il faut parfois refaire certaines parties du modèle de données dans l'application web. Ce manque de souplesse est en partie dû à la technologie utilisée. Les SGBD tel que PostgreSQL sont stables et performants mais manquent de souplesse face à d'autres technologies comme des SGBD no SQL qui supportent plus facilement les changements de structure dans les données.

LA GÉNÉRICITÉ ET LES META-DONNEES

L'une des problématiques principales pour l'intégration de données fut la modélisation du schéma relationnel de la base de données. Ce schéma doit être construit de manière à pouvoir stocker les données phénotypiques issues de plusieurs espèces, collectées sur des dispositifs hétérogènes à plusieurs échelles sur des sites multiples, et impliquant plusieurs partenaires et plusieurs projets.

DA PHNE
DAtabase PHenotype plaNt intEgration

A PROPOS CONTACT PROJET ÉQUIPE

MEDHI

Import de données

Formulaire d'importation de données vers DAPHNE

Import de données

Formulaire de données

Aucun formulaire sélectionné

- Public uniquement
- Dispositif expérimental
- Accessions
 - Associer des accessions à des unités expérimentales
 - Échantillons
- Accès contrôlés
 - Itinéraires techniques

Visibilité

Public

Importer

Figure 17 Interface d'importation et de saisie des données : choix du formulaire de données

Ce sont les conditions auxquelles la base de données doit répondre pour être un outil générique de capitalisation de ces données.

Pour pouvoir comparer les données de plusieurs espèces il faut pouvoir utiliser un vocabulaire commun. Pour cela la structure particulière des tables qui accueillent les données phénotypiques permettent le stockage de ces données sous n'importe quel nom de variable définit au préalable par l'utilisateur. Dans une base de données standard il faudrait rajouter une nouvelle colonne pour chaque nouvelle variable définie dans DAPHNE et donc modifier la structure de la base de données. Ces opérations consomment énormément de ressources et la plupart des SGBD ne sont pas optimisés pour ce genre de pratique.

C'est pourquoi l'informaticienne de l'équipe (Mme. Sandrine Auzoux) a suggéré d'utiliser une méthode déjà mise en œuvre dans le projet ECOFI à savoir, la création d'une table référençant les variables et de construire les tables qui accueillent les données de manière à ce que chaque entrée de ces tables soient en relation avec une entrée de la table qui référence les variables. De cette manière on associe les valeurs à leur variable ligne par ligne et non rangées par colonne.

CONCLUSION ET PERSPECTIVES

Pour conclure, DAPHNE permet la gestion des données issues des essais agronomiques. Ce système d'information est conçu en prévoyant l'interopérabilité avec d'autres systèmes en charge de la gestion des ressources génétiques et des ontologies agronomiques. Il permet de centraliser et de pérenniser les données phénotypiques accumulées depuis plusieurs années de recherche et répartis dans divers fichiers. Mais aussi, aux agronomes d'analyser, valider et comparer leurs résultats de recherche sur plusieurs années. L'application web, encore en développement, permet actuellement de définir le design des dispositifs expérimentaux, de tracer les interventions faites au champ et de définir les groupes et les projets dans lesquels les utilisateurs évoluent. L'intégration de ses données se fait par l'intermédiaire de formulaires de saisie ou par l'importation de fichier CSV spécifiques et dont les patrons sont téléchargeables depuis l'application.

Cependant, Il reste encore plusieurs aspects à développer pour ce projet : (i) Etablir les liens avec les systèmes existant de gestion des ressources génétiques et des ontologies agronomiques. (ii) Ajouter des formulaires à l'interface web pour pouvoir intégrer la totalité des données phénotypiques recueillies par le CIRAD. (iii) Développer des interfaces pertinentes pour l'interrogation et la visualisation des données. (iv) Effectuer des tests unitaires pour assurer la performance et la qualité du traitement des données et des requêtes.

Durant cette formation, j'ai pu mettre à profit mes connaissances et acquérir de nouvelles compétences en développement de base de données et en programmation web au travers l'utilisation de Frameworks. Ce stage m'a permis de découvrir les activités bio-informatiques d'un centre de recherche agronomique. J'ai pu constater l'importance d'être à l'interface entre la biologie et l'informatique dans un contexte d'intégration des données et le support d'une équipe pluridisciplinaire que requiert la réalisation d'un tel projet. Aussi, je me suis très bien intégré à l'équipe ID en participant à leur réunion hebdomadaire. Cette équipe dynamique a été un support important à la réalisation de l'application web. Par ailleurs, j'ai eu le plaisir de participer à la présentation de nos travaux à la réunion annuelle du projet BFF de l'INRA à Versailles et à la création d'un poster scientifique sur DAPHNE [Annexe : 4]. Les encadrants de cette formation étant satisfaits des résultats obtenus, m'ont proposé un contrat de 3 mois afin de continuer le développement de la base de données et de l'interface web. Et je souhaite vivement continuer cette expérience avec eux.

REFERENCES

- Auzoux Sandrine, R. L.-F. (2016). ECOFI: A generic agronomic database to facilitate analysis and crop modelling. *Crop modelling for agriculture and food security under global change*.
- C. Pommier, E. K. (2009). Ephesis: Environment and Phenotypes Information System, a GnpIS module. *TDWG 2009 Annual conference*. Montpellier, France.
- Ken Schwaber, M. B. (2001). *Agile Software Development with Scrum*. Prentice Hall.
- Reenskaug, T. (1979). *The original MVC reports*. Dept. of Informatics University of Oslo.
- Schneier, B. (1993). Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish). *Springer-Verlag*.
- The Integrated Breeding Platform. (s.d.). *The Breeding Management System Version 3.0.8* (2015). Récupéré sur <https://www.integratedbreeding.net/breeding-management-system>

ANNEXES

ANNEXE 1 : LES OUTILS DE DÉVELOPPEMENTS

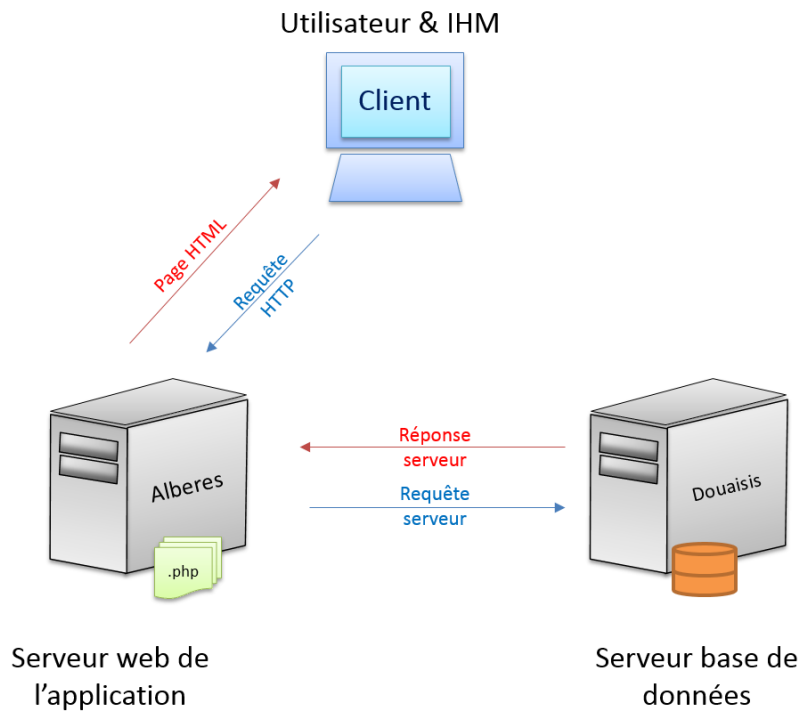
LA BASE DE DONNEES

- **PgAdmin III** : Plateforme d'administration et de développement libre pour PostgreSQL la plus populaire.
- **PhpPgAdmin** : Interface web d'administration.
- **PostgreSQL Maestro©** : Outil propriétaire complet pour l'administration et le développement de base de données PostgreSQL. Il permet aussi de faire de la visualisation. L'interface graphique est ergonomique et facilite la modélisation du schéma relationnel.
- **SQL Power Architect** : Logiciel de visualisation UML de base de données.

L'APPLICATION WEB

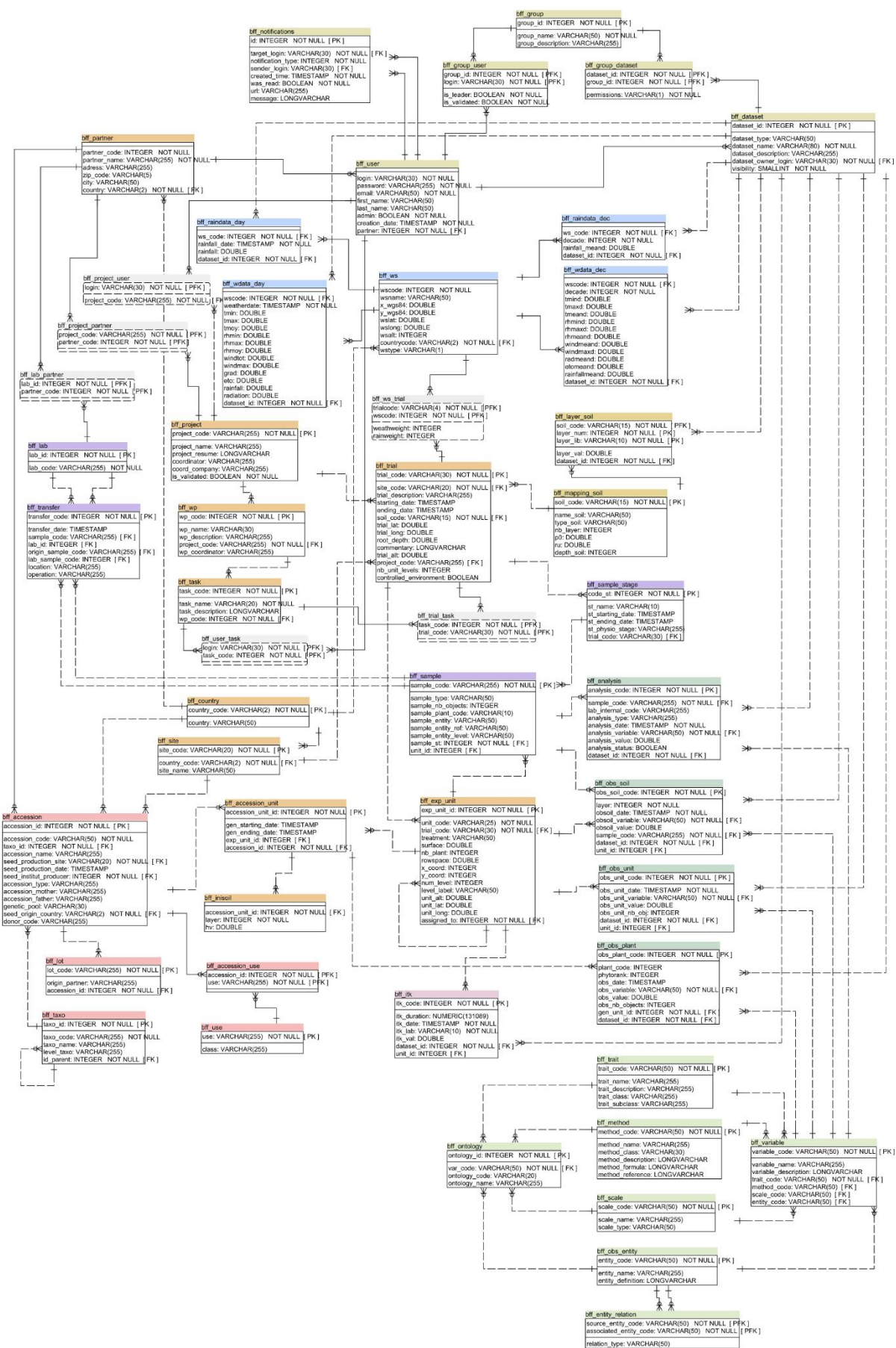
- **CodeIgniter 3** : Framework PHP pour le développement d'application web
- **Bootstrap** : Framework front-end pour la gestion des éléments graphique d'une interface web.
- **Atom** : L'IDE (Environnement de développement) développé par l'équipe de GitHub. (<https://atom.io/>)

ANNEXE 2 : L'ARCHITECTURE 3 TIERS



- Niveau 1 : L'utilisateur demande la ressource web et le navigateur est en charge d'afficher les données transmises par le serveur web.
- Niveau 2 : Le serveur qui héberge l'application est en charge du traitement des données provenant du navigateur et de la base de données DAPHNE hébergée sur un autre serveur. Sur la plateforme SouthGreen qui gère l'infrastructure, le serveur hébergeant l'application web se nomme « **Alberes** » (Nom de domaine : alberes.CIRAD.fr).
- Niveau 3 : Le serveur PostgreSQL est en charge du SGBD, c'est donc lui qui stocke les données et qui les envoie vers le serveur web de l'application. Ce serveur PostgreSQL a pour nom « **Douaisis** » (nom de domaine : douaisis.CIRAD.fr).

ANNEXE 3 : SCHEMA UML DE DAPHNE



Résumé : Mon stage de fin d'étude s'inscrit dans le cadre du projet Biomass For the Future (BFF) portant sur la thématique de l'agroécologie. Les objectifs de ce stage sont la construction d'une base de données pour stocker, sécuriser et capitaliser les données issues des différents domaines suivant : écophysiologie, météorologie, génétique, histologie, biochimie ; et le développement d'une application web permettant l'alimentation et l'interrogation de cette base de données. La première partie de ce manuscrit est dédiée à l'étude des solutions existantes pour l'intégration des données phénomiques en agronomie et la présentation des objectifs de ce stage. La seconde partie concerne les outils et les méthodologies mises en œuvre pour le développement de DAPHNE (DAtabase PHenotype plaNt intEgration), ainsi que la modélisation de cette base de données. Enfin, la dernière partie est consacrée aux différents sous-systèmes de la base de données et les fonctionnalités développées dans l'application web. DAPHNE a été conçue grâce aux systèmes de gestion de données PostgreSQL et son interface est réalisée en langage PHP à l'aide de la structure logicielle CodeIgniter.